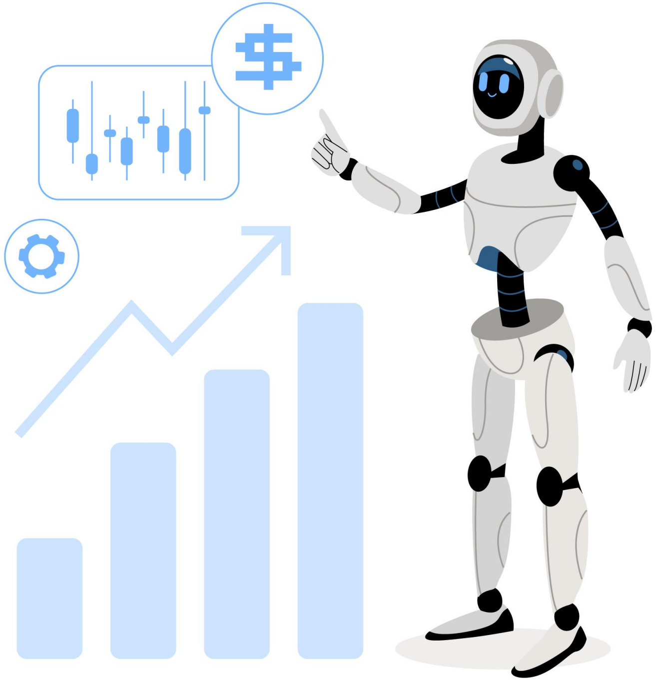
IIT
INSTITUT INTERNATIONAL TECHNOLOGIE
Université Nord Américaine Privée

Data Mining Project Presentation

Insurance Claim Prediction (Classification Problem)

Presented by:

Mahassen Drira
Chedy Chaaben
Taieb Jemal



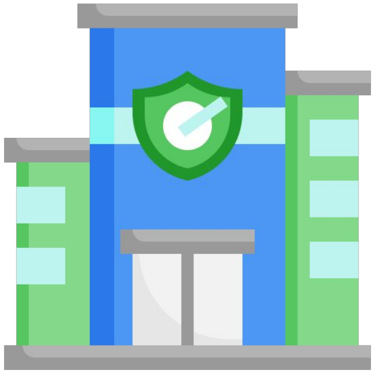
1 Project Context

2 Objectives

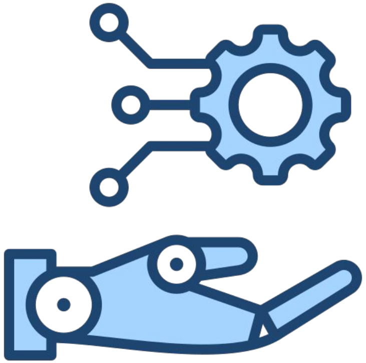
3 Proposed Solution

4 Conclusion

Project Context

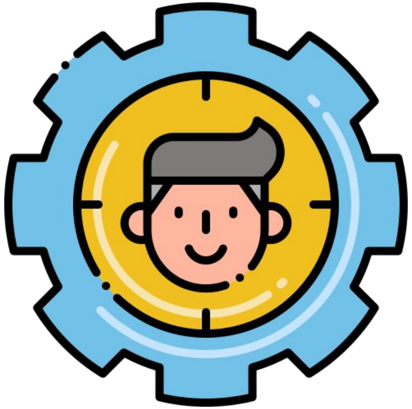


Insurance Company



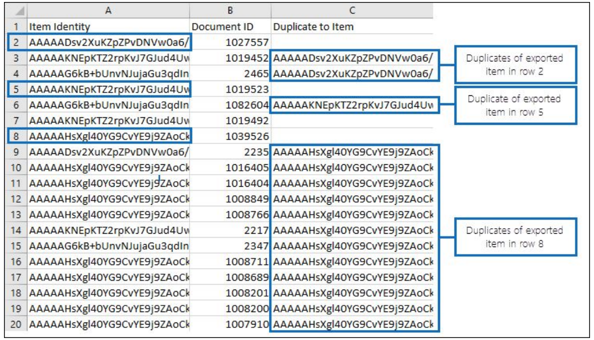
Process Automation

Objective



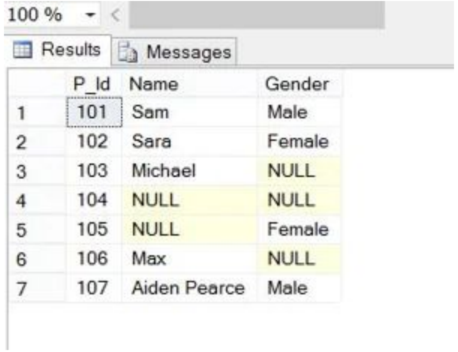
Personalized and high-quality service

Problems



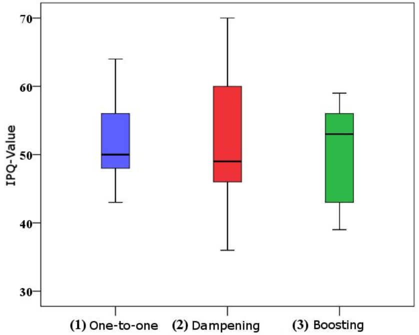
	A	B	C
1	Item Identity	Document ID	Duplicate to Item
2	AAAAADsv2XuKZpZPvDNVw0a6/	1027557	
3	AAAAAKNEpKTZ2rpKvJ7GJud4Uw	1019452	AAAAADsv2XuKZpZPvDNVw0a6/
4	AAAAAG6kB+bUnvNJujaGu3qdIn	2465	AAAAADsv2XuKZpZPvDNVw0a6/
5	AAAAAKNEpKTZ2rpKvJ7GJud4Uw	1019523	
6	AAAAAG6kB+bUnvNJujaGu3qdIn	1082604	AAAAAKNEpKTZ2rpKvJ7GJud4Uw
7	AAAAAKNEpKTZ2rpKvJ7GJud4Uw	1019492	
8	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1039526	
9	AAAAADsv2XuKZpZPvDNVw0a6/	2235	AAAAAHsXgl40YG9CvYE9j9ZAoCk
10	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1016405	AAAAAHsXgl40YG9CvYE9j9ZAoCk
11	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1016404	AAAAAHsXgl40YG9CvYE9j9ZAoCk
12	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008849	AAAAAHsXgl40YG9CvYE9j9ZAoCk
13	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008766	AAAAAHsXgl40YG9CvYE9j9ZAoCk
14	AAAAAKNEpKTZ2rpKvJ7GJud4Uw	2217	AAAAAHsXgl40YG9CvYE9j9ZAoCk
15	AAAAAG6kB+bUnvNJujaGu3qdIn	2347	AAAAAHsXgl40YG9CvYE9j9ZAoCk
16	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008711	AAAAAHsXgl40YG9CvYE9j9ZAoCk
17	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008689	AAAAAHsXgl40YG9CvYE9j9ZAoCk
18	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008201	AAAAAHsXgl40YG9CvYE9j9ZAoCk
19	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1008200	AAAAAHsXgl40YG9CvYE9j9ZAoCk
20	AAAAAHsXgl40YG9CvYE9j9ZAoCk	1007910	AAAAAHsXgl40YG9CvYE9j9ZAoCk

Duplicated Rows

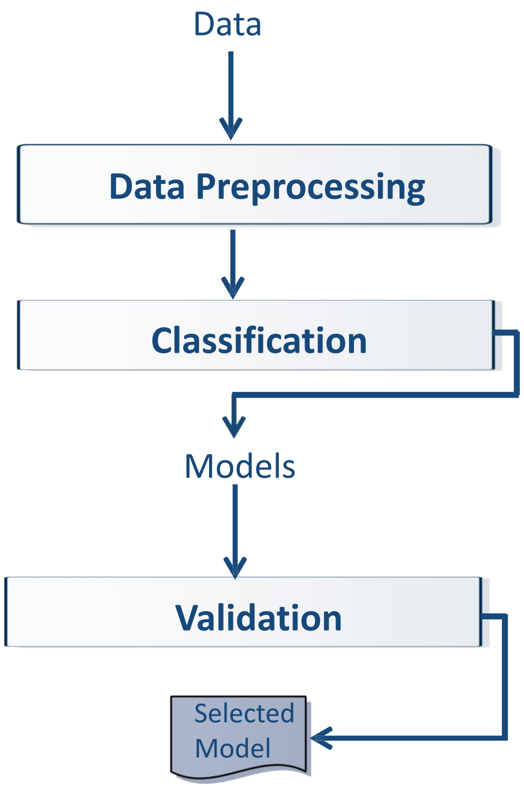


	P_Id	Name	Gender
1	101	Sam	Male
2	102	Sara	Female
3	103	Michael	NULL
4	104	NULL	NULL
5	105	NULL	Female
6	106	Max	NULL
7	107	Aiden Pearce	Male

Null Values



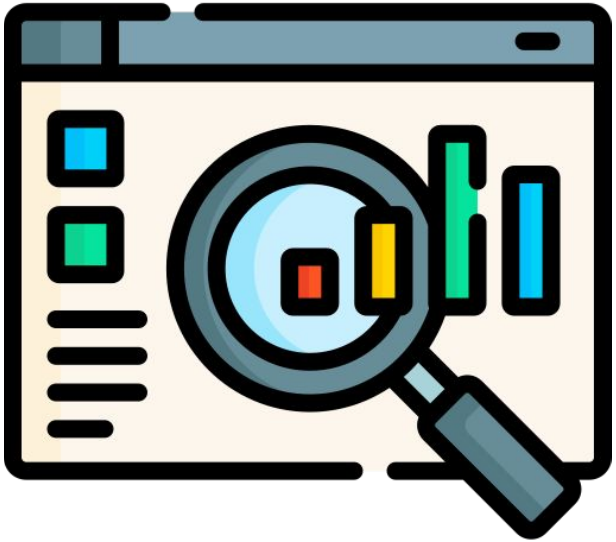
Outliers Presence



Knowledge Discovery from Data (ECD)

1. Data Preprocessing:

Inspect the data:



1. Data Preprocessing:

Analyze :

	Customer Id	YearOfObservation	Insured_Period	Residential	Building_Painted	Building_Fenced	Garden	Settlement	Building Dimension	Building_Type	NumberOfWindows	Geo_Code	Claim
0	H13501	2012	1.0	1	N	V	V	U	1240.0	Wood-framed	without	75117	non
1	H14962	2012	1.0	0	N	V	V	U	900.0	Non-combustible	without	62916	non
2	H17755	2013	1.0	1	V	N	O	R	4984.0	Non-combustible	4	31149	oui
3	H13369	2016	0.5	0	N	V	V	U	600.0	Wood-framed	without	6012	oui
4	H12988	2012	1.0	0	N	V	V	U	900.0	Non-combustible	without	57631	non

1.Data Preparation :

- Analyze :

	Customer Id	YearOfObservation	Insured_Period	Residential	Building_Painted	Building_Fenced	Garden	Settlement	Building Dimension	Building_Type	NumberOfWindows	Geo_Code	Claim
count	5012	5012.000000	5012.000000	5012.000000	5012	5012	5008	5012	4935.000000	5012	5012	4939	5012
unique	5012	NaN	NaN	NaN	2	2	2	2	NaN	4	11	1115	2
top	H13501	NaN	NaN	NaN	V	N	O	R	NaN	Non-combustible	without	6088	non
freq	1	NaN	NaN	NaN	3763	2535	2532	2537	NaN	2310	2476	102	3886
mean	NaN	2013.660215	0.869713	0.301077	NaN	NaN	NaN	NaN	1876.898683	NaN	NaN	NaN	NaN
std	NaN	1.383134	0.219496	0.458772	NaN	NaN	NaN	NaN	2267.277397	NaN	NaN	NaN	NaN
min	NaN	2012.000000	0.500000	0.000000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN	NaN	NaN
25%	NaN	2012.000000	0.500000	0.000000	NaN	NaN	NaN	NaN	520.000000	NaN	NaN	NaN	NaN
50%	NaN	2013.000000	1.000000	0.000000	NaN	NaN	NaN	NaN	1067.000000	NaN	NaN	NaN	NaN
75%	NaN	2015.000000	1.000000	1.000000	NaN	NaN	NaN	NaN	2280.000000	NaN	NaN	NaN	NaN
max	NaN	2016.000000	1.000000	1.000000	NaN	NaN	NaN	NaN	2 20840.000000	NaN	NaN	NaN	NaN

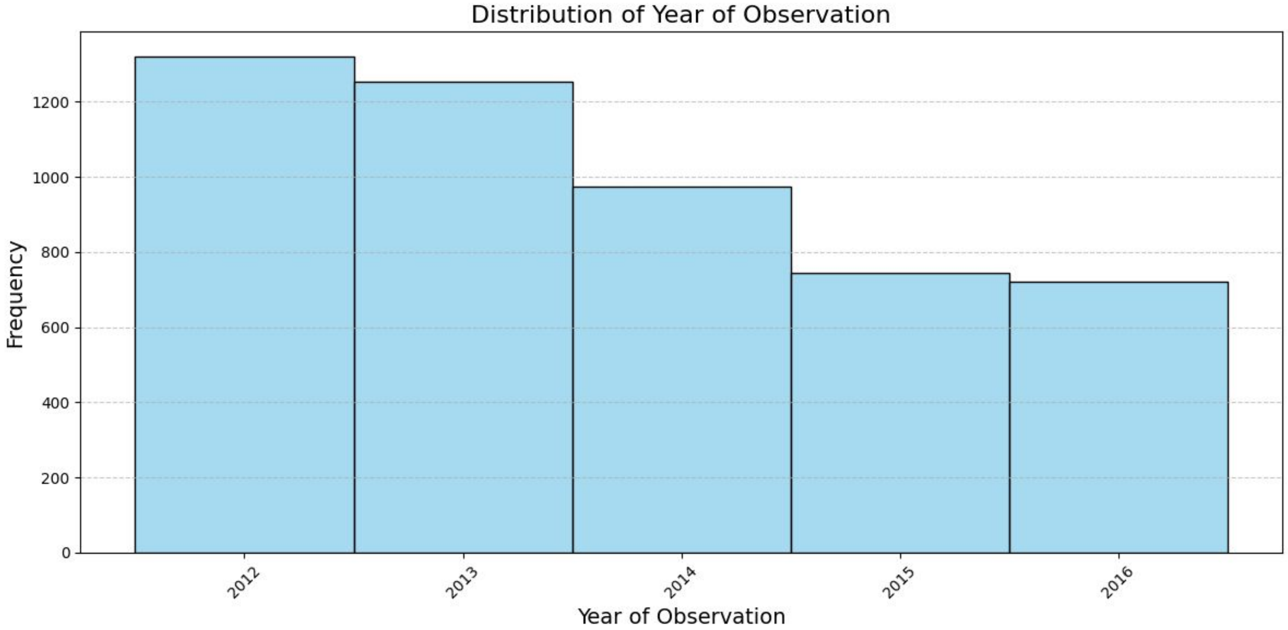
1. Data Preprocessing:

- Analyze :

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	Customer Id	5012 non-null	object
1	YearOfObservation	5012 non-null	int64
2	Insured_Period	5012 non-null	float64
3	Residential	5012 non-null	int64
4	Building_Painted	5012 non-null	object
5	Building_Fenced	5012 non-null	object
6	Garden	5008 non-null	object
7	Settlement	5012 non-null	object
8	Building Dimension	4935 non-null	float64
9	Building_Type	5012 non-null	object
10	NumberOfWindows	5012 non-null	object
11	Geo_Code	4939 non-null	object
12	Claim	5012 non-null	object

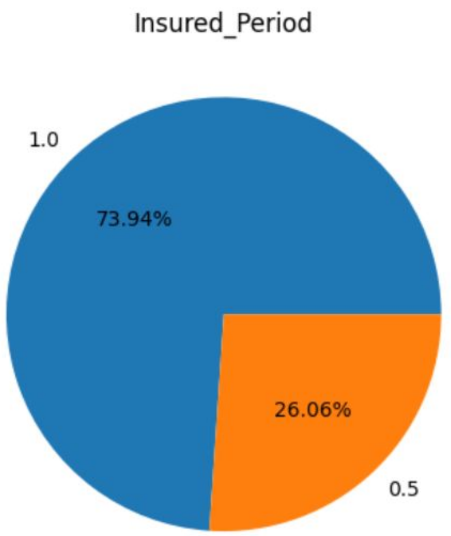
1. Data Preprocessing:

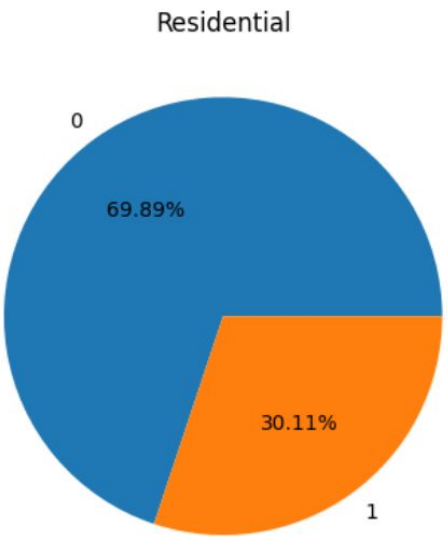
- Visualize :

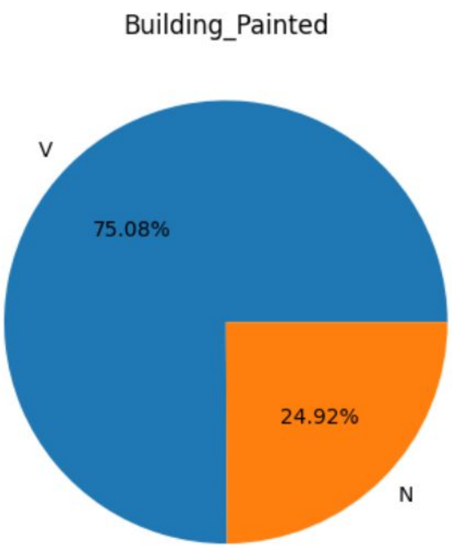


1. Data Preprocessing:

- Visualize :

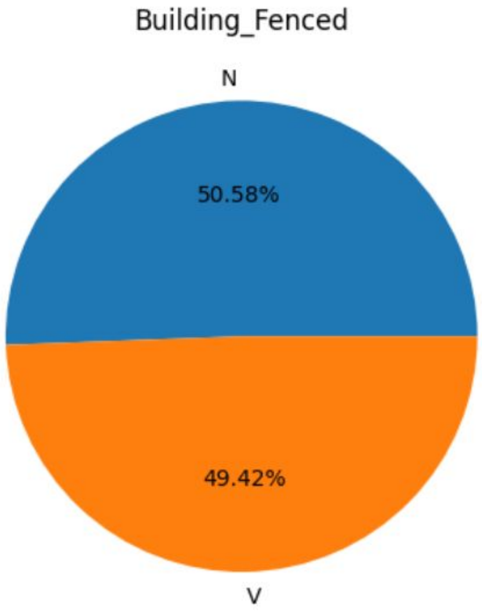


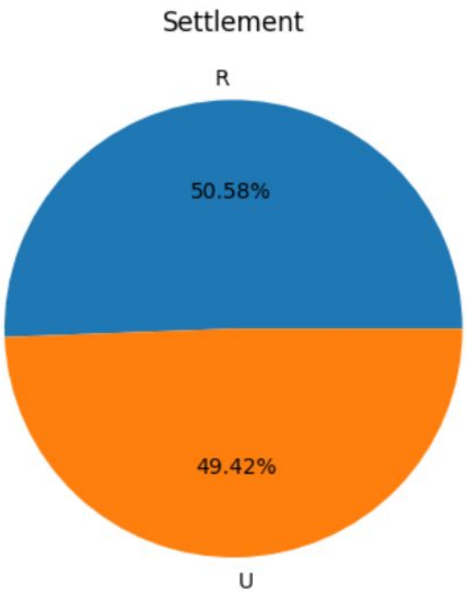




1. Data Preprocessing:

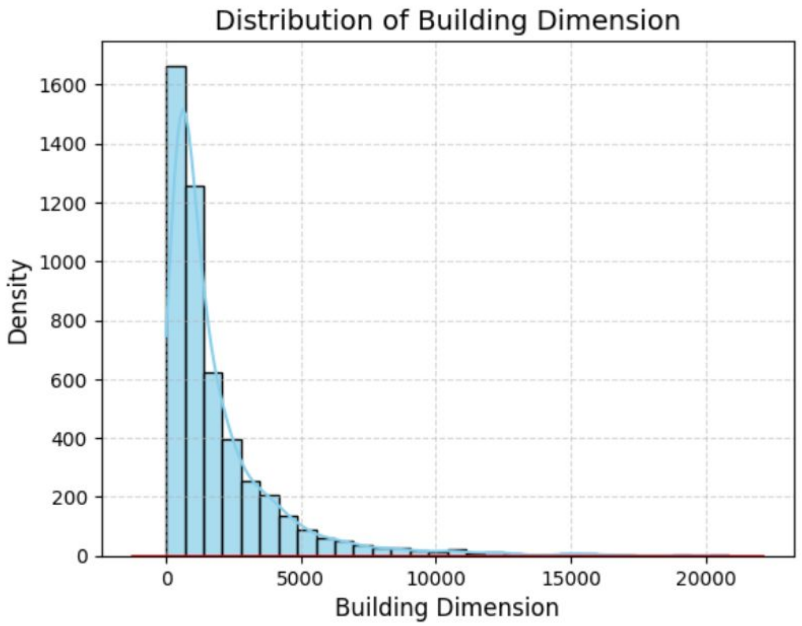
- Visualize :

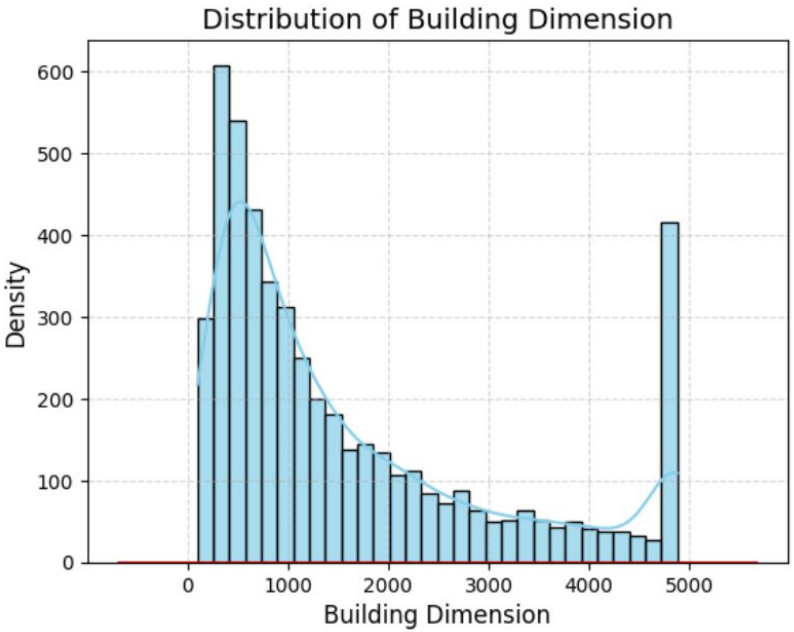




1. Data Preprocessing:

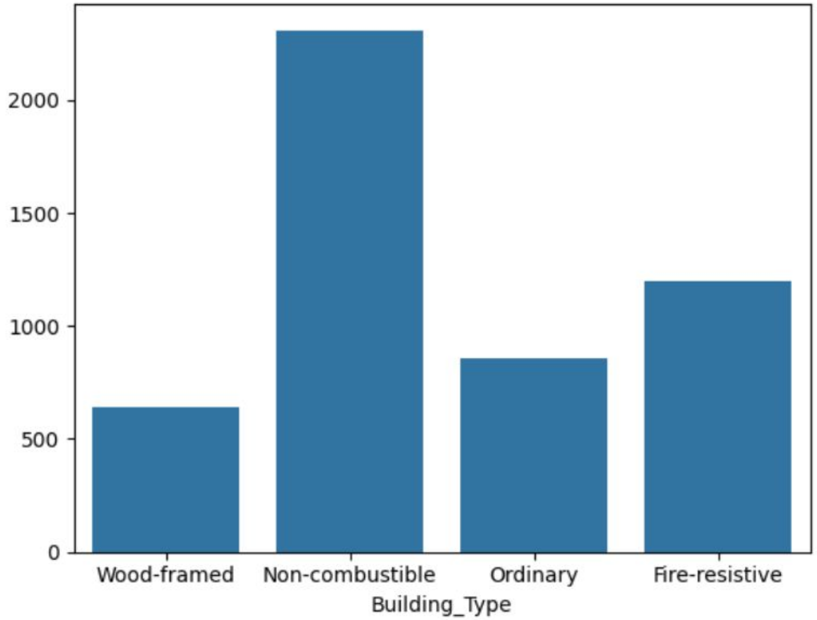
- Visualize (Building Dimension):





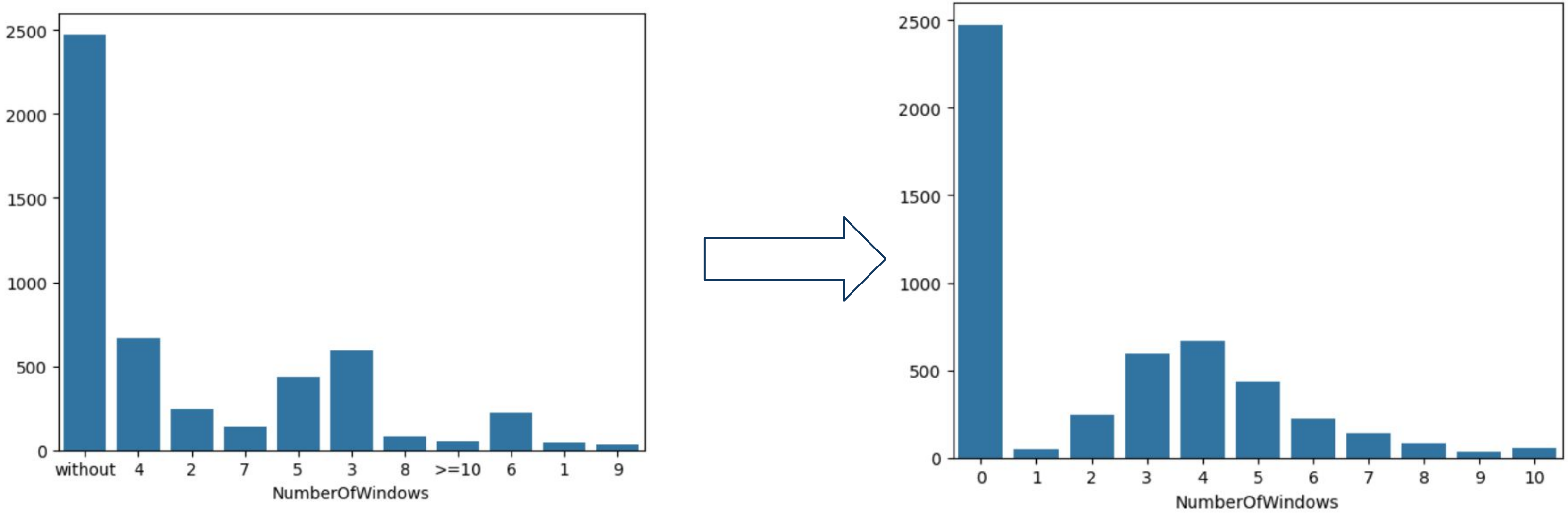
1. Data Preprocessing:

- Visualize (Building Type):



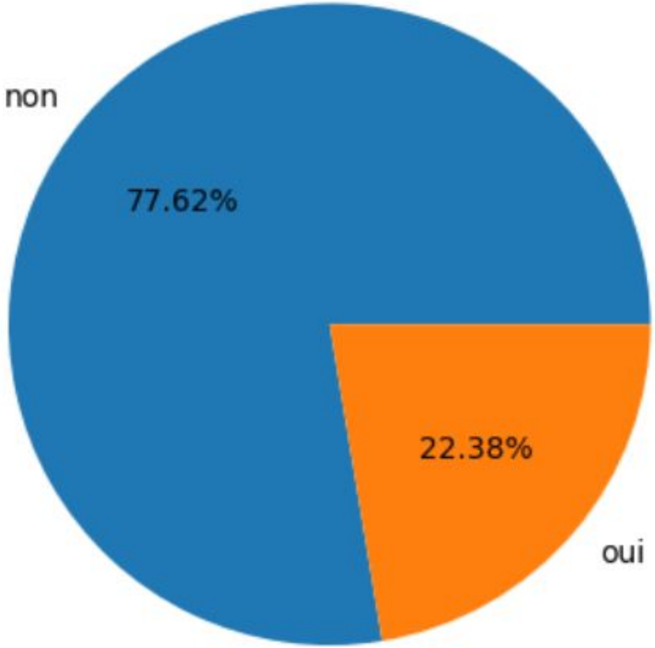
1. Data Preprocessing:

- Clean & Visualize (Number Of Windows):



1. Data Preprocessing:

- Visualize (Claim):



Data Preprocessing:

Clean Duplicated Rows → Make Data Ready For Classification → Visualize

2. Dimension reduction:

(Useless Features)

- Year Of Observation
- Customer Id

Nomber of None Values = 0

	Customer Id
count	5012
unique	5012
top	H13501
freq	1

Nomber of None Values = 0

	YearOfObservation
count	5012.000000
mean	2013.660215
std	1.383134
min	2012.000000
25%	2012.000000
50%	2013.000000
75%	2015.000000
max	2016.000000

```
Index(['Insured_Period', 'Residential', 'Building_Painted', 'Building_Fenced',
       'Garden', 'Settlement', 'Building Dimension', 'Building_Type',
       'NumberOfWindows', 'Geo_Code', 'Claim'],
      dtype='object')
```

2. Binary Encoding :

- Building_Painted : (N : oui, V : non) - - > (1 : oui, 0 : non)
- Building_Fenced : (N : oui, V : non) - - > (1 : oui, 0 : non)
- Garden : (V : oui, O : non) - - > (1 : oui, 0 : non)
- Settlement - - > urbain_zone

  (R : zone rurale, U : zone urbain) - - > (1 : zone urbain , 0 : zone rurale)

- Number Of Windows

  (without dans le cas de 0 fenêtre) - - > (0 dans le cas de 0 fenêtre)

- Claim : (oui : Claim , non : Not Claim) - - > (1 : Claim , 0 : Not Claim)

2. Missing Value Handling:

- Garden (dropna)
- Building Dimension (Simple Imputer(most_frequent))
- Geo_Code (forward fill)
- State + City Density (dropna)

```
# Fonction pour le traitement des valeur manquantes
def traitement_des_valeurs_manquantes(df,NomDuColone):
  mf_imputer = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
  df[NomDuColone] = mf_imputer.fit_transform(df[[NomDuColone]])
  return df
```

2. Outliers:

- Building Dimension

```
# Fonction pourt elimination des outliers
def treatment_des_outliers(df,feature):
  Q1,Q3=np.percentile(df[feature],[25,75])
  IQR=Q3-Q1
  lower_limit=max(Q1 - 1.5 * IQR, df[feature].min()+100)
  # Lower_limit is -2125 building dimension can t be negatif nor close to 0
  upper_limit=Q3+1.5*IQR
  df[feature]=np.where(df[feature]>=upper_limit,
  upper_limit, np.where(df[feature]<=lower_limit,
  lower_limit,df[feature]))
  return df
```

2.Discretization:

Q1(33%) : 650.0
Q2(66%) : 1699.2400000000007

```
count     5008.000000
mean      1611.475040
std       1428.627826
min        101.000000
25%        500.000000
50%       1037.500000
75%       2250.000000
max       4875.000000
Name: Building Dimension, dtype: float64
```

Building Dimension
3760.0
1452.0
1944.0
2270.0
2976.0
...
862.0
NaN
730.0
568.0
730.0

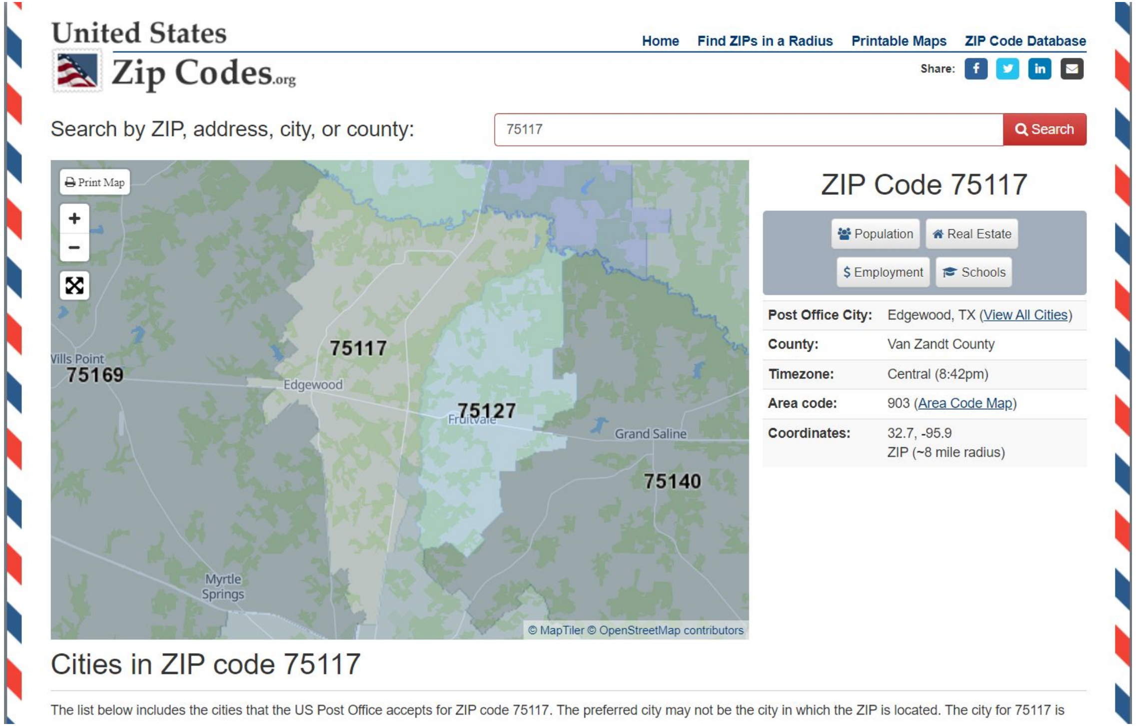
Small_Building	Medium_Building	Large_Building
0	0	1
0	1	0
0	0	1
0	0	1
0	0	1
...	...	...
0	1	0
1	0	0
0	1	0
1	0	0
0	1	0

2. One Hot Encoder:

Building_Type
Fire-resistive
Fire-resistive
Ordinary
Non-combustible
Fire-resistive
...
Wood-framed
Non-combustible
Non-combustible
Non-combustible
Non-combustible

Building_Type_Fire-resistive	Building_Type_Non-combustible	Building_Type_Ordinary	Building_Type_Wood-framed
1	0	0	0
1	0	0	0
0	0	1	0
0	1	0	0
1	0	0	0
...	...	...	...
0	0	0	1
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0

2. Geo_Code :


United States
Zip Codes.org
Home
Find ZIPs in a Radius
Printable Maps
ZIP Code Database
Share:
Search by ZIP, address, city, or county:
75117
Search
Print Map
75117
Vills Point
75169
Edgewood
75127
Fruitvale
Grand Saline
75140
Myrtle
Springs
© MapTiler © OpenStreetMap contributors
ZIP Code 75117
Population
Real Estate
$ Employment
Schools
Post Office City: Edgewood, TX (View All Cities)
County: Van Zandt County
Timezone: Central (8:42pm)
Area code: 903 (Area Code Map)
Coordinates: 32.7, -95.9
ZIP (~8 mile radius)
Cities in ZIP code 75117
The list below includes the cities that the US Post Office accepts for ZIP code 75117. The preferred city may not be the city in which the ZIP is located. The city for 75117 is

2. Geo_Code :

simplemaps
Interactive Maps & Data

Databases	Basic	Pro	Comprehensive
Commercial use	Allowed	Allowed	Allowed
File format	CSV, Excel	CSV, Excel, SQL	CSV, Excel, SQL
Census-designated zips	Yes, all ZCTAS	Yes, all ZCTAS	Yes, all ZCTAS
Current USPS zips	Most	Yes, all USPS zips	Yes, all USPS zips
Number of entries	33,783	41,561	41,561
Fields (listed below)	Basic fields	More fields	All fields
Future updates	Not guaranteed	Included for 12 months	Included for 24 months
Attribution	Required	Not required	Not required
License	Creative Commons Attribution 4.0	Permissive, no redistribution	Permissive, no redistribution
Refund policy	N/A	30-day guarantee	30-day guarantee
One-time fee	Free	$99	$199
	Download	Buy Now!	Buy Now!

https://simplemaps.com/data/us-zips

2. Geo_Code :

	state_id	zip	density
0	PR	601	100.2
1	PR	602	477.6
2	PR	603	543.1
3	PR	606	47.3
4	PR	610	264.4
..	...	...	...
101	PR	911	6028.4
102	PR	912	6474.9
103	PR	913	7984.8
104	PR	915	6743.9
105	PR	917	5151.5

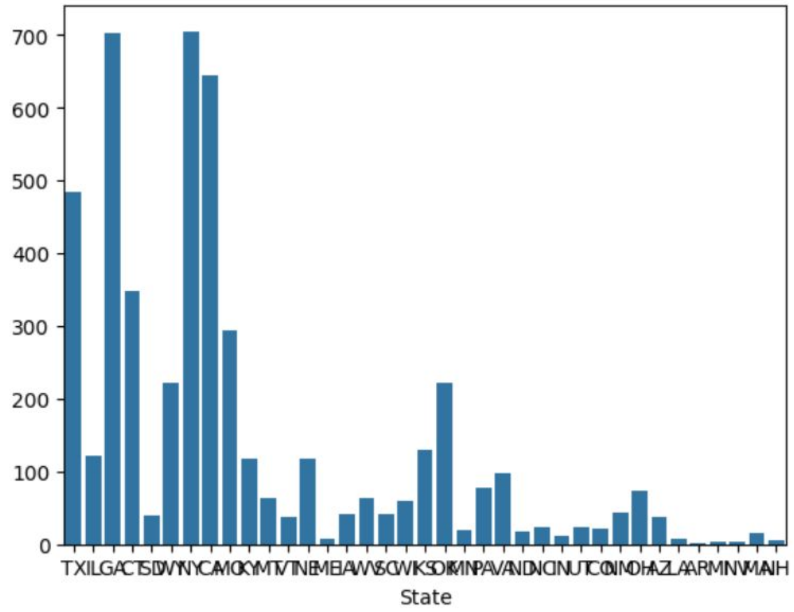
	state_id	zip_start	zip_end	new_density
0	PR	601	987	1105.897710
1	MA	1001	2791	1218.233581
2	RI	2802	2921	1148.051852
3	NH	3031	3897	123.766802
4	ME	3901	4992	67.660798
5	VT	5001	5907	90.939623
6	CT	6001	6907	646.406597
7	NY	6390	14905	2141.604605
8	NJ	7001	8904	1532.152843
9	PA	15001	19611	533.255950
10	DE	19701	19980	564.182353
11	DC	20001	20591	3083.259649
12	VA	20105	24657	378.867996
13	MD	20601	21930	617.602516
14	WV	24701	26886	76.365718
15	NC	27006	28909	240.443611

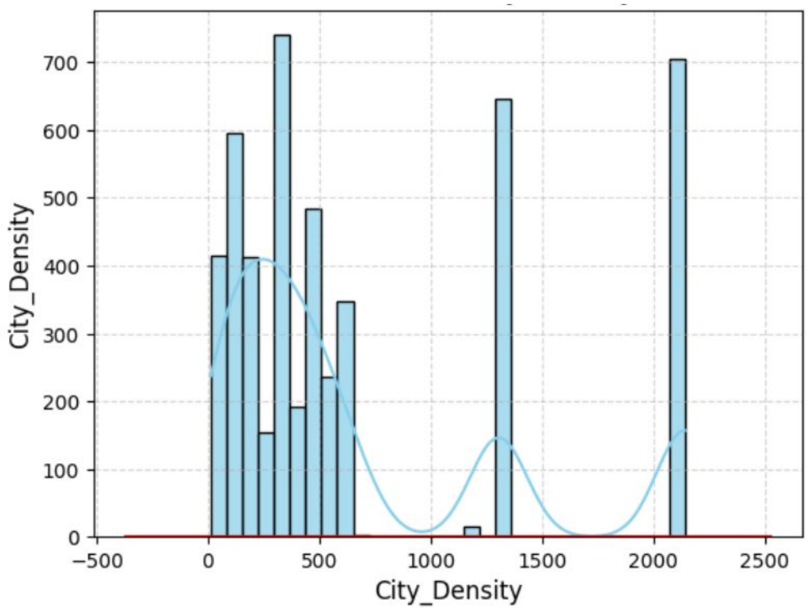
2. Geo_Code :

Building_Type_Ordinary	Building_Type_Wood-framed	NumberOfWindows	State	City_Density	Claim
0	0	0	OH	374.595377	oui
0	0	5	ND	52.396649	non
1	0	6	CA	1306.829079	oui
0	0	0	NY	2141.604605	oui
0	0	9	VT	90.939623	non
...	...	...	...	...	...
0	1	2	CT	646.406597	non
0	0	0	CT	646.406597	non
0	0	3	NE	126.082765	non
0	0	0	NE	126.082765	oui
0	0	0	CA	1306.829079	non

2. Geo_Code :

- Visualize (Geo Code - > State + City Density):





2. Missing Value Handling:

- State + City Density (dropna)

```
Nomber of None Values =  65
```

	State
count	4943
unique	37
top	NY
freq	705

dtype: object

Label Encoding State

- State

State
OH
ND
CA
NY
VT
...

State
25
19
2
24
33
...

Converting Float to Int for city Density with astype

- City_Density

City_Density
374.595377
52.396649
1306.829079
2141.604605
90.939623
...

City_Density
374
52
1306
2141
90
...

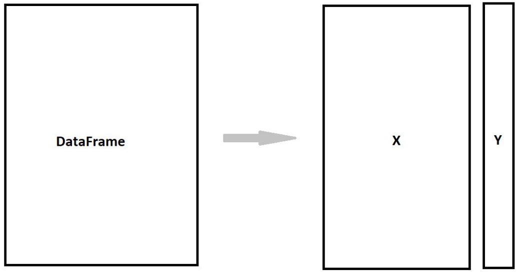
Removed Duplicate Lines

	Insured_Period	Residential	Building_Painted	Building_Fenced	Garden	urbain_zone	Small_Building	Medium_Building	Large_Building
58	1.0	1	[illegible]	0	1	1	0	1	0
87	1.0	0	0	0	1	1	1	0	0
91	1.0	1	1	0	1	1	[illegible]	1	0
103	1.0	[illegible]	1	0	1	1	[illegible]	1	0
127	1.0	[illegible]	1	[illegible]	1	1	[illegible]	0	0
...	...	[illegible]	...	...	...	...	...	...	...
4937	0.5	[illegible]	1	0	[illegible]	1	[illegible]	0	0
4938	1.0	[illegible]	1	0	1	1	[illegible]	0	0
4939	0.5	0	0	1	0	0	[illegible]	1	0
4940	1.0	1	0	1	0	[illegible]	1	0	0
4941	0.5	0	0	1	0	0	1	0	0

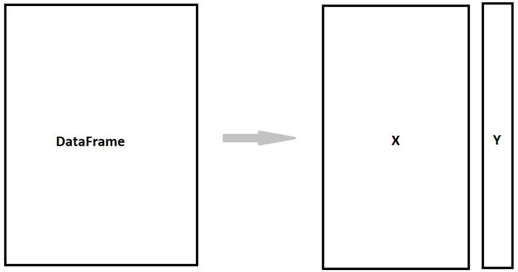
2819 rows × 17 columns

2. Pre-Classification :

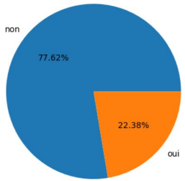
X_train, Y_train



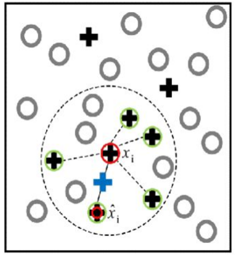
X_test, Y_test



Unbalanced Data
Problem (Claim)



SMOTE



2. Classification:

DecisionTreeClassifier

DecisionTreeClassifier(max_depth=2)

SVC

SVC(probability=True)

GradientBoostingClassifier

GradientBoostingClassifier()

MLPClassifier

MLPClassifier(max_iter=500)

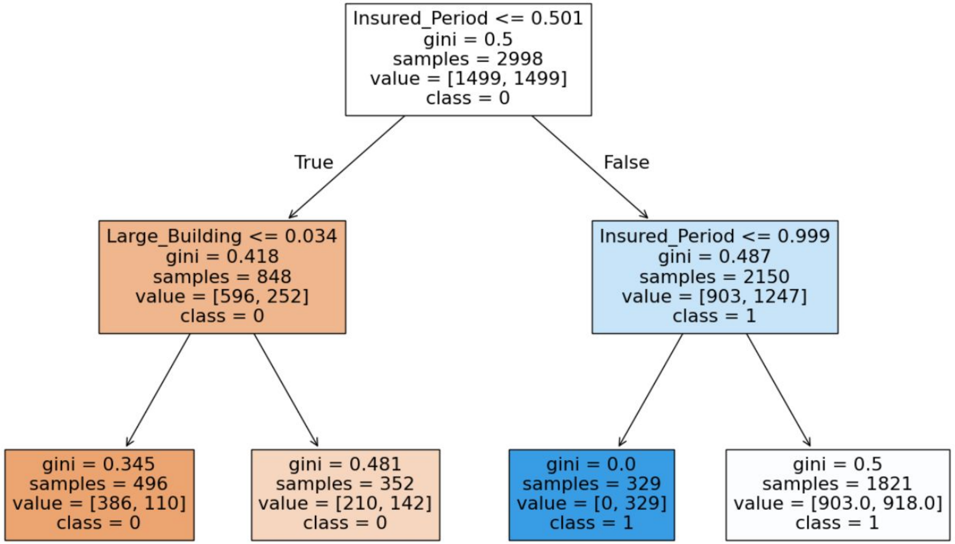
RandomForestClassifier

RandomForestClassifier()

3.Models :

Decision Tree (Visualisation):

Plot Tree (Figure)



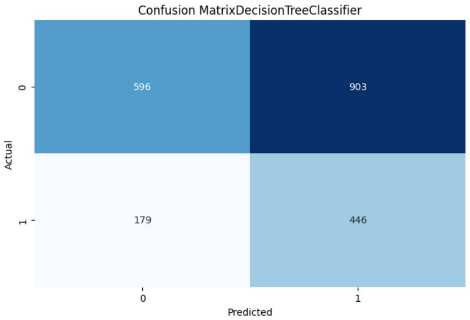
Plot Tree (Exported Text)

```
|--- Insured_Period <= 0.50
|   |--- Large_Building <= 0.03
|   |   |--- class: 0
|   |--- Large_Building >  0.03
|   |   |--- class: 0
|--- Insured_Period >  0.50
|   |--- Insured_Period <= 1.00
|   |   |--- class: 1
|   |--- Insured_Period >  1.00
|   |   |--- class: 1
```

3. Models :

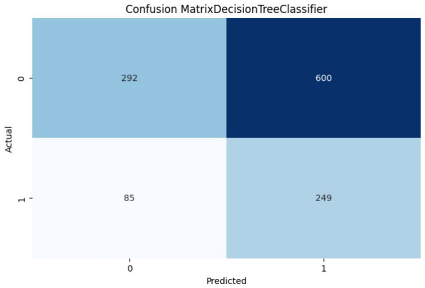
Decision Tree (Evaluation):

Train Set




```
'Train Set Evaluation For DecisionTreeClassifier'
Accuracy: 0.49
F1 Score: 0.45
Area Under Curve: 0.56
              precision    recall  f1-score   support

           0       0.77      0.40      0.52      1499
           1       0.33      0.71      0.45       625

    accuracy                           0.49      2124
   macro avg       0.55      0.56      0.49      2124
weighted avg       0.64      0.49      0.50      2124
```

Test Set

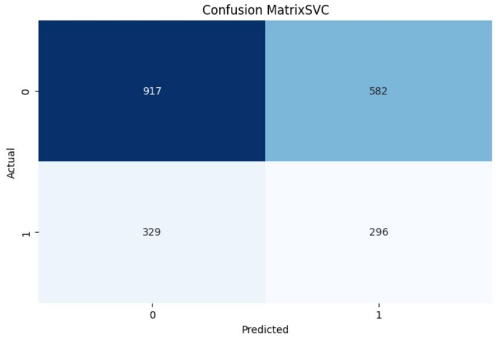



```
'Test Set Evaluation ForDecisionTreeClassifier'
Accuracy: 0.44
F1 Score: 0.42
Area Under Curve: 0.55
              precision    recall  f1-score   support

           0       0.77      0.33      0.46       892
           1       0.29      0.75      0.42       334

    accuracy                           0.44      1226
   macro avg       0.53      0.54      0.44      1226
weighted avg       0.64      0.44      0.45      1226
```

3. Models :

SVC (Evaluation) :

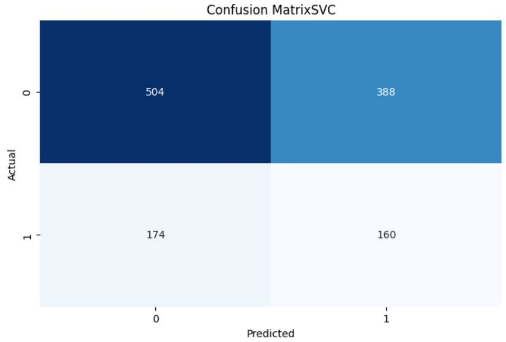
Train Set



'Train Set Evaluation For SVC'
Accuracy: 0.57
F1 Score: 0.39
Area Under Curve: 0.56

	precision	recall	f1-score	support
0	0.74	0.61	0.67	1499
1	0.34	0.47	0.39	625
accuracy			0.57	2124
macro avg	0.54	0.54	0.53	2124
weighted avg	0.62	0.57	0.59	2124

Test Set



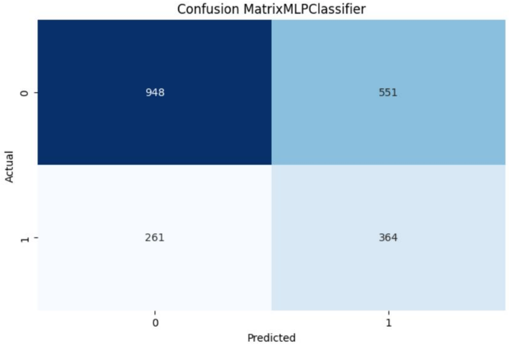
'Test Set Evaluation ForSVC'
Accuracy: 0.54
F1 Score: 0.36
Area Under Curve: 0.54

	precision	recall	f1-score	support
0	0.74	0.57	0.64	892
1	0.29	0.48	0.36	334
accuracy			0.54	1226
macro avg	0.52	0.52	0.50	1226
weighted avg	0.62	0.54	0.57	1226

3. Models :

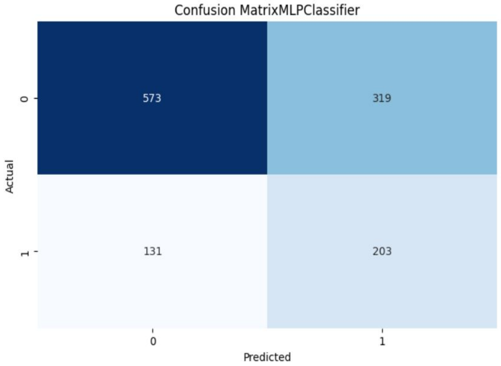
MLP Classifier :

Train Set




```
'Train Set Evaluation For MLPClassifier'
Accuracy: 0.62
F1 Score: 0.47
Area Under Curve: 0.64
              precision    recall  f1-score   support

           0       0.78      0.63      0.70      1499
           1       0.40      0.58      0.47       625

    accuracy                           0.62      2124
   macro avg       0.59      0.61      0.59      2124
weighted avg       0.67      0.62      0.63      2124
```

Test Set

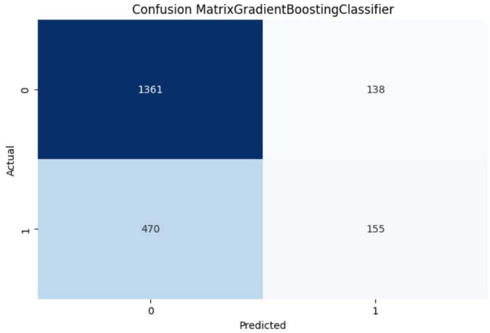



```
'Test Set Evaluation ForMLPClassifier'
Accuracy: 0.63
F1 Score: 0.47
Area Under Curve: 0.66
              precision    recall  f1-score   support

           0       0.81      0.64      0.72       892
           1       0.39      0.61      0.47       334

    accuracy                           0.63      1226
   macro avg       0.60      0.63      0.60      1226
weighted avg       0.70      0.63      0.65      1226
```

3. Models :

Gradient Tree Boosting :

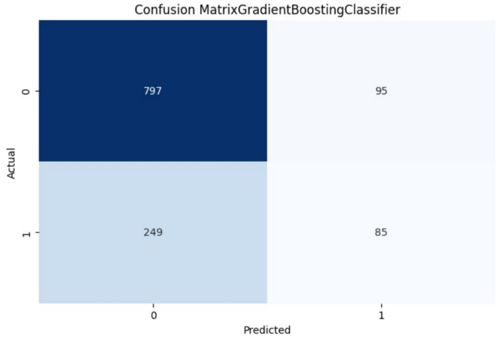
Train set



'Train Set Evaluation For GradientBoostingClassifier'
Accuracy: 0.71
F1 Score: 0.34
Area Under Curve: 0.69

	precision	recall	f1-score	support
0	0.74	0.91	0.82	1499
1	0.53	0.25	0.34	625
accuracy			0.71	2124
macro avg	0.64	0.58	0.58	2124
weighted avg	0.68	0.71	0.68	2124

Test set



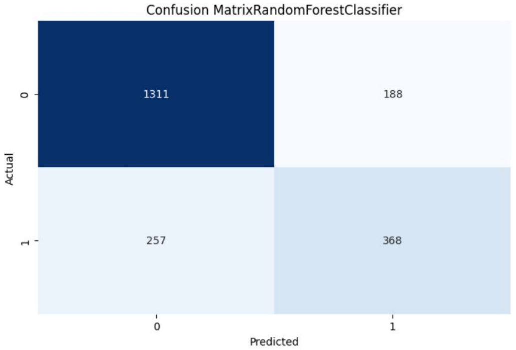
'Test Set Evaluation ForGradientBoostingClassifier'
Accuracy: 0.72
F1 Score: 0.33
Area Under Curve: 0.67

	precision	recall	f1-score	support
0	0.76	0.89	0.82	892
1	0.47	0.25	0.33	334
accuracy			0.72	1226
macro avg	0.62	0.57	0.58	1226
weighted avg	0.68	0.72	0.69	1226

3. Models :

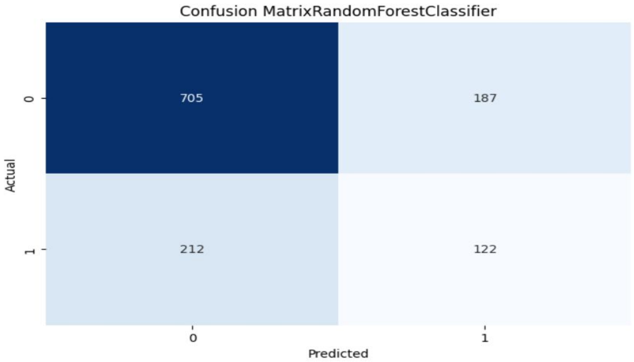
RandomForestClassifier :

Train set




```
'Train Set Evaluation For RandomForestClassifier'
Accuracy: 0.79
F1 Score: 0.62
Area Under Curve: 0.89
              precision    recall  f1-score   support

           0       0.84      0.87      0.85      1499
           1       0.66      0.59      0.62       625

    accuracy                           0.79      2124
   macro avg       0.75      0.73      0.74      2124
weighted avg       0.78      0.79      0.79      2124
```

Test set




```
'Test Set Evaluation ForRandomForestClassifier'
Accuracy: 0.67
F1 Score: 0.38
Area Under Curve: 0.63
              precision    recall  f1-score   support

           0       0.77      0.79      0.78       892
           1       0.39      0.37      0.38       334

    accuracy                           0.67      1226
   macro avg       0.58      0.58      0.58      1226
weighted avg       0.67      0.67      0.67      1226
```

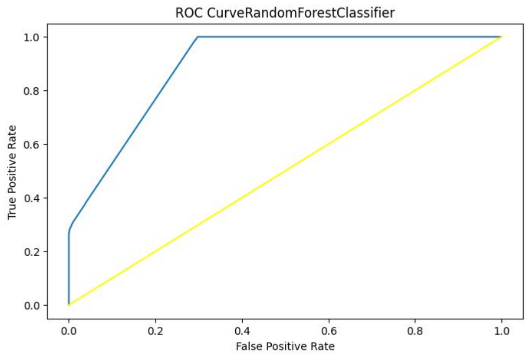
4. Validation :

RandomForestClassifier

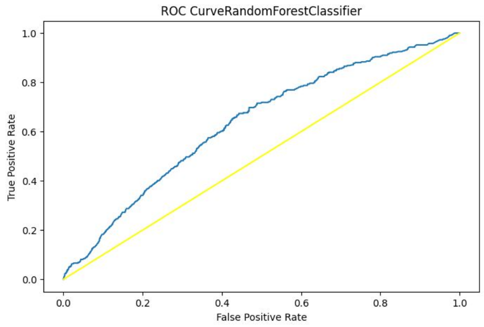
Thanks to its ability to capture complex relationships between variables, it makes the most of the available information.
It offers:

Acceptable accuracy > 70%
Acceptable area under curve 1 < AUC < 0.5

Train Set



Test Set



Areas of Improvement: Selecting better HyperParameters with the help of Random Search CV OR Grid Search CV

Conclusion

We achieved optimal performance through effective data preprocessing, model training, and selection.

Thank you

INSTITUT INTERNATIONAL TECHNOLOGIE
Université Nord Américaine Privée